DOCUMENT RESUME

ED 204 405 TM 810 424

AUTHOR Kolen, Michael J.; Whitney, Douglas R.
TITLE Comparison of Four Procedures for Equating the Tests of General Educational Development.
PUB DATE Apr 81
NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; *Equated Scores; *Equivalency Tests; *Factor Analysis; *Latent Trait Theory; Testing Problems
IDENTIFIERS *Cross Validation; Equipercentile Equating; *General Educational Development Tests; Linear Equating Method; Rasch Model; Three Parameter Model

ABSTRACT

Procedures used to compare the results from item response theory as well as more traditional equating methods were described and critically analyzed. The implications of the comparison of equipercentile, linear, one-parameter (Rasch), and three-parameter methods for equating twelve forms of each of the five tests of General Educational Development (GED) were discussed. The use of factor analyses to assess test dimensionality, examination of equating curves, examination of item parameter estimates for extremes, comparison of equating sample means and variances, and cross-validation analyses were recommended for use by testing programs contemplating a switch from traditional to item response theory equating. The three-parameter equating method produced unacceptable equating results--possibly because only 200 examinees per equating form were used. The one-parameter (Rasch) method produced results which were as stable as those for the traditional methods. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED204405

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

COMPARISON OF FOUR PROCEDURES FOR EQUATING THE TESTS OF GENERAL EDUCATIONAL DEVELOPMENT

Michael J. Kolen
Hofstra University

Douglas R. Whitney
American Council on Education

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
M. J. Kolen
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the
American Educational Research Association
in Los Angeles, California, April 1981.

TM 810124

COMPARISON OF FOUR PROCEDURES FOR EQUATING THE TESTS OF GENERAL EDUCATIONAL DEVELOPMENT

Michael J. Kolen
Hofstra University

Douglas R. Whitney
American Council of Education

ABSTRACT

Procedures used to compare the results from item response theory as well as more traditional equating methods were described and critically analyzed. The implications of the comparison of equipercentile, linear, one-parameter (Rasch), and three-parameter methods for equating twelve forms of each of the five tests of General Educational Development (GED) were discussed.

The use of factor analyses to assess test dimensionality, examination of equating curves, examination of item parameter estimates for extremes, comparison of equating sample means and variances, and cross-validation analyses were recommended for use by testing programs contemplating a switch from traditional to item response theory equating. The three-parameter equating method produced unacceptable equating results--possibly because only 200 examinees per equating form were used. The one-parameter (Rasch) method produced results which were as stable as those for the traditional methods.

Comparison of Four Procesures for Equating the Tests
of General Educational Development

Most testing programs construct comparable forms of their tests for a variety of reasons such as maintaining test security and enabling an individual to take a test more than once. Appropriate test score equating procedures allow the forms to be used interchangeably without serious question as to the comparability of common scale scores. Appropriate test score equating enables us to say that, "A score of 50 means the same thing, whether it is earned on form 1 or form 2." While equipercentile and linear equating methods (Angoff, 1971) are the most widely accepted procedures for equating tests, item response theory methods have recently been advocated as being more flexible (Lord, 1980 and Wright and Stone, 1980).

Many testing programs might want to convert from equipercentile or linear equating to item response theory methods in order to gain this additional flexibility. Although the conversion probably could be justified only if it were not accompanied by a loss in equating accuracy or precision, few systematic and objective procedures for comparing the adequacy of results from two or more equating methods exist. Thus, the practitioner has no prespecified set of systematic and objective procedures to aid in choosing among competing methods.

The major purpose of the present study was to describe and critically analyze the procedures used to compare the results from item response as well as more traditional equating methods, as applied to the equating of twelve forms of each of the five tests of General Educational Development (GED). The procedures included the use of cross-validation

criteria and factor analyses were used to assess the degree of unidimensionality of the GED tests.

Another purpose was to present the results from an empirical comparison of equipercentile, linear, one-parameter (Rasch) logistic item response theory, and three-parameter logistic item response theory methods for equating the GED tests. The GED tests are achievement tests administered to approximately one-half million adults each year; qualified individuals earn a high school equivalency diploma or certificate. The implications of the findings for using item response theory methods to equate the GED as well as other tests were discussed.

The test forms were administered to examinees in pairs (Design II - Angoff, 1971). Although the procedures used to compare equating methods can be used with any equating design, most of them require the existence and use of a representative cross-validation group which has taken the two equated forms. Also discussed was a procedure which can be used when randomly equivalent groups of cross-validation examinees have taken the two equated forms. The results from the equating methods were compared on the raw score scale.

Equating Requirements

Lord (1980) stated three requirements for methods of equating two unidimensional tests. The first was referred to as equity. Assume that a group of individuals of exactly the same ability has been identified and that each individual has taken the same pair of equated test forms. Also, assume that the scores on the second form have been converted to the score scale of the first form using the equating results. For equity to hold, the distribution of scores on the first form must be

identical to the distribution of converted scores on the second form. This property must hold for a group of individuals at any given ability. Note that the equity implies that the distribution of scores on the two forms, after conversion to a common score scale, must be identical for any group of individuals, regardless of the distribution of ability in the group.

The second property was referred to as invariance across groups. For invariance to hold, the equating results must be the same, regardless of the group of individuals used to equate the tests.

The third property is symmetry. That is, the equating should be the same regardless of which test is equated to the scale of the other. This rules out, for example, linear regression as an equating method.

Lord (1980) showed that equating of observed scores can be expected to meet the equity and invariance requirements only when the tests to be equated are either identical, in which case equating would not be needed anyway, or perfectly reliable, a condition which will not occur in practice. It seems reasonable however, that equating methods which come closest to meeting the equity and invariance requirements for observed scores should be preferred.

A few empirical studies which examined the equity and invariance requirements have been completed. These are reviewed in Kolen (1981). In general, the results are inconclusive as to which equating methods are to be preferred in most practical situations. The relative degree to which the requirement of equity was met was the primary procedure used to compare the equating methods in the present study.

Two assumptions are required when item response theory methods are used to scale tests. First, the test must be unidimensional, or

alternatively, the items must exhibit the property of local independence (Lord and Novick, 1968). Although no completely satisfactory procedures for assessing test dimensionality exist, Lord (1980) suggested using results from a factor analysis of the inter-item tetrachoric correlation matrix. This procedure was used in the present study.

The second assumption is that the item response curves follow the prespecified functional form. This assumption was addressed indirectly in the present study by comparing the equating results from the item response and traditional equating methods with respect to the equity requirement.

The GED Tests

The GED tests are used to evaluate learning in everyday life, enabling qualified individuals to earn high school equivalency diplomas or certificates. Through the GED Testing Service of the American Council on Education, the tests were administered to nearly one-half million candidates in 1979. According to the GED Teacher's Manual (1979, p. 5), "The GED tests are designed to measure, as nearly as possible, the major and lasting outcomes and skills generally associated with four years of regular high school instruction."

The GED consists of tests in each of five subject matter areas. The Writing Skills test (80 items) contains items in spelling, capitalization and punctuation, usage, sentence correction, and logic and organization. The Social Studies test (60 items) contain U.S. history, economics, geography, political science, and behavioral science items. The Science test (60 items) contains items from biology, earth sciences, chemistry, and physics. The fourth test, Reading Skills (40 items), contains prac-

tical reading, general reading, prose literature, poetry, and drama items. The Mathematics test (50 items) contains arithmetic, geometry, and algebra items. The items on the Mathematics test, for the most part, are story problems rather than straight computational items.

Blocks of items, where two or more items relate to a common stimulus, are included in substantial numbers on the Social Studies, Science, and Reading Skills tests. The common stimuli consist of passages, graphs, charts, etc. Approximately one-third of the Social Studies items, two-thirds of the Science items, and almost all of the Reading Skills items are contained in common stimulus blocks. A more detailed description of the tests is presented in the GED Teacher's Manual (1979). Because of the apparent content heterogeneity of the GED tests and the inclusion of many common stimulus blocks, assessment of the degree of unidimensionality of each test prior to the item response theory analyses seemed very desirable.

Test Development

Twelve forms of each of the five GED tests were developed and standardized by the Educational Testing Service (ETS) between December 1974 and January 1978. The tests were constructed in conjunction with subject-matter advisory panels. In Spring 1977, the current GED tests were standardized and equated using a carefully selected, stratified random sample of high school students in the United States. Kuder-Richardson 20 reliabilities of the forms ranged from .84 to .95 across the five tests. A variety of validity studies were also completed. These studies and a more complete description of the GED development and standardization are provided in ETS (1978).

1977 GED Equating Sample

The 1977 GED equating sample data were used in the present study; for this reason they will be described in detail. The design for collecting the data involved randomly sampling 294 school districts stratified by public-private, geographic region, and socio-economic status, from among U.S. school districts. One high school was randomly sampled from each district and 22 students were to be sampled from each school for use in the equating portion of the studies.

The twelve GED test forms included in the present study will be referred to as the anchor form and equating forms one through eleven. Each examinee was administered the anchor form and one randomly selected equating form of two of the GED tests. The order in which the forms were administered was counterbalanced. These procedures resulted in 2227, 2278, 2267, 2269, and 2244 usable anchor form/equating form pairs for the Writing Skills, Social Studies, Science, Reading Skills, and Mathematics tests, respectively. Approximately 205 examinees were administered each equating form of each of the tests.

Procedure

Principal axis factor analyses were completed to examine the unidimensionality assumption. Tetrachoric correlation matrices for each test were factored using the squared multiple correlation of an item with all other items as communalities. The degree of unidimensionality exhibited by each test was assessed through examination of the eigenvalues.

Twelve forms of each of the five GED tests were separately equated using equipercentile and linear equating methods as well as one-parameter (Rasch) and three-parameter logistic estimated true score equivalents

equating methods. The item parameter estimates were examined for extreme values, the equating curves were studied, and the equating results were compared using a cross-validation sample. The data source and the procedures used for the equating and cross-validation comprise the remainder of this section.

Data Source

The 1977 GED equating sample was used as the data source. Note that examinees were administered the anchor form and one of the equating forms of a test. Whenever an examinee correctly answered either all or none of the items on the anchor form or the equating form, the examinee's data for that test were removed from the present study. This procedure was followed because item response theory estimation procedures cannot estimate the ability of individuals earning all or none correct on a form. Between 190 and 218 examinees took each test and equating form combination.

Two hundred examinee records were then randomly selected, stratified by geographic region and socio-economic status, from each test and equating form combination. These examinees comprised the cross-validation sample and were not used in the equating portion of the study. The remaining 170 to 198 examinees per test and equating form combination will be referred to as the equating sample.

Equating Methodology

Four equating methods were used to equate the GED forms using the equating sample data. Linear and equipercentile methods are discussed together and referred to as traditional equating methods. One-parameter logistic (Rasch) and three-parameter logistic methods are discussed together and referred to as item response theory equating methods.

Traditional equating. The anchor and equating form pairs of each test were equated separately. For example, the 198 equating sample examinees taking both the anchor form and equating form one of the Writing Skills test were used to equate form one to the anchor form raw score scale. Method IA-1 described by Angoff (1971) was used for linear equating and Method IA-2 was used for equipercentile equating.

For linear equating, whenever the anchor form equivalent of an equating form one score was above the highest possible score on the anchor form, it was fixed at the highest possible anchor form score. A similar procedure was followed whenever the anchor form equivalent was below a score of zero. For equipercentile equating, linear interpolation, as opposed to smoothing, was used when necessary. Identical procedures were followed in the equating of equating forms one through eleven scores to the anchor form raw score scale for each of the five GED tests.

Item response theory methods. The first step in item response theory equating was to estimate the item and ability parameters for the one-parameter and three-parameter logistic models. The LOGIST computer program of Wood, Wingersky, and Lord (1976) was used for this purpose.

The anchor form and equating form one of the Writing Skills test will be used as an example. The item parameters for the 80 anchor form items and the ability parameters for the 2,227 equating sample examinees who took the Writing Skills test were estimated using LOGIST. The ability parameters for the 198 examinees who also were administered equating form one were then fixed. These fixed ability estimates along with the item responses of these 198 examinees were then entered into LOGIST. Because of small sample sizes, the "pseudo-chance" parameters for the equating forms were fixed at the modal anchor form value of the corre-

ponding anchor test.[1] This produced equating form one item parameters on the same scale as the anchor form estimates. Similar procedures were followed for equating forms two through eleven of the Writing Skills test. These procedures were also followed for the other four GED tests using both the one-parameter and three-parameter logistic item response theory models.[2]

The next stage in the equating process was to derive anchor form score equivalents of equating form scores using estimated true score equating (Lord, 1980). The estimated true score of an examinee with a given estimated ability is equal to the sum, over items, of the estimated probability of correctly answering each item. Using non-linear estimation procedures, anchor form estimated true score equivalents of equating form one through eleven integer scores were calculated. The procedure was followed for the five GED tests using the one-parameter and three-parameter logistic models.

Note that estimated true scores below the estimated "pseudo-chance" level of a test (the sum of the item "pseudo-chance" parameter estimates) are undefined for the three-parameter logistic model. Scores of zero on any pair of forms were arbitrarily considered to be equivalent; "missing"

---

[1]The modal "pseudo-chance" level parameters were 0.150, 0.165, 0.140, 0.200, and 0.150 for the GED Writing Skills, Social Studies, Science, Reading Skills, and Mathematics tests, respectively.

[2]An attempt was made to simultaneously estimate all of the item parameters on the 12 forms of the Writing Skills test using the three-parameter logistic model. LOGIST failed to converge, however. Lord (1980, pp. 209-210) suggested a modification to the LOGIST program which could be expected to solve the convergence problem. The simultaneous procedure also could be expected to produce more precise item parameter estimates because the ability estimates would reflect performance on all items taken rather than the anchor form items only. The authors were unaware of Lord's modification at the time this study was conducted.

equivalents below the "pseudo-chance" were arrived at via linear interpolation. Lord (1980, pp. 210-211) addressed this problem in a slightly different manner.

Cross-Validation Methodology

The twenty randomly selected examinees from each test and equating form combination comprised the cross-validation sample. The anchor form and equating form one scores on the Writing Skills test will be used as an example in the development of the cross-validation procedures.

The twenty cross-validation sample examinee scores on equating form one were converted to the anchor form score scale using the linear method equating table. Let X_i represent the score of cross-validation sample examinee i on the anchor form of the Writing Skills test. Let Y_i represent the score of the same examinee equating form one and let Y_i' represent this equating form one score converted to the anchor form score scale using the linear method equating table for converting equating form one scores to the anchor form scale. The difference between the anchor form score (X_i) and the converted equating form one score (Y_i') for an examinee,

$$D_i = X_i - Y_i', \tag{1}$$

was used as the basis for forming a cross-validation summary statistic. The quantity D_i reflects both equating error and errors of measurement.

The D_i quantities could be squared and then averaged over the twenty appropriate cross-validation examinees. However, this quantity can be broken down into further components.

$$\frac{\Sigma D_i^2}{n} = \bar{D}^2 + \frac{\Sigma(D_i - \bar{D})^2}{n}. \tag{2}$$

In this equation n is the number of cross-validation examinees (20 for the present study). The first quantity to the right of the equal sign is the mean value of D, squared. This quantity represents the squared, mean difference between anchor form and converted equating form one scores. It will be referred to as the measure of equating bias. The second quantity on the right represents the variance of the differences between anchor form and converted equating form one scores and will be referred to as the measure of equating imprecision.

Equating bias and imprecision indices were computed separately for each test and equating form combination. A one-way repeated measures analysis of variance was completed for each test and index combination. Form (eleven levels--equating forms one through eleven) was treated as the random "subjects" factor and equating method (four levels) as the fixed repeated measures factor. Tukey post-hoc paired comparisons were also used.

Kolen (1981) developed a cross-validation index which is appropriate when randomly equivalent groups take the forms to be equated. The index can also be applied when each examinee takes both tests, such as in the present study. This index will be referred to as the percentile comparison index.

The percentile comparison index is a measure of the dissimilarity between distributions of anchor form scores and converted scores on an equating form. To compute this index, the cross-validation distributions were tabulated and percentile ranks calculated separately for the anchor form and converted equating form one scores. The percentile comparison index was formed by finding the difference between each observed anchor form score and the converted equating form one score with an identical

percentile rank in the converted equating form one distribution. This difference was then weighted by the number of individuals earning the anchor form score and summed over the observed anchor form scores. The equation is,

$$\frac{\Sigma f_i (X_i - Y_i'')^2}{n} \qquad (3)$$

In the equation, X_i represents an anchor form integer score, Y_i'' the equating form one score with the same percentile rank, and f_i the number of examinees that earned X_i. Like the bias and imprecison indices, smaller values indicate better performance for the equating method. Repeated measures analyses of variance were completed for the percentile comparison index in a manner similar to those completed for the bias and imprecision measures.

Results

The eigenvalues and percentages of variance accounted for by each of the first twelve factors in the factor analyses are presented in Table 1. The ratios of the first to second eigenvalue, a rough index of

---

Insert Table 1 About Here

---

unidimensionality, were 7.4, 10.2, 7.4, 9.3, and 4.7 for the Writing Skills, Social Studies, Science, Reading Skills, and Mathematics tests, respectively. Only the Mathematics test approached having a substantial second factor. Overall, the factor analyses suggested that all of the tests, except possibly Mathematics, were reasonably unidimensional.

The item parameter estimates were examined for irregularities. Extreme three-parameter model difficulty estimates (absolute value above 3.5) were discovered for a number of items on the equating forms of the Writing Skills, Social Studies, and Science tests. Very few were found on the Reading Skills and Mathematics tests. These are reflected in the standard deviations of the three-parameter item difficulty parameter estimates shown in Table 2.

Insert Table 2 About Here

Note that extreme three-parameter discrimination estimates were not reflected in the standard deviations in Table 2. However, the discrimination estimates were constrained between 0 and 2 by LOGIST. Also, the discrimination index may not be on an equal-interval scale; differences in parameter estimates near zero may reflect larger differences in discrimination than differences at other points. The fact that very low discrimination estimates tended to accompany the extreme difficulty estimates supports this notion. The one-parameter model produced no extreme difficulty estimates across all of the tests and forms. It appears that problems were encountered in estimating the item parameters in the three-parameter model, especially for the longer tests.

The equating relationships were also examined. Figure 1 presents the equating relationships between the anchor form and equating form one of the Reading Skills test. This pair of forms was chosen because equating form one contained no extreme parameter estimates and the relationships were fairly representative.

Insert Figure 1 About Here

Note that the anchor form was generally less difficult than equating form one. This was true for most forms studied. The mean anchor form raw scores were generally from one to three points higher than their equating form counterparts across all tests. This result is illustrated in Figure 1.

In the figure, the three-parameter method produced the smallest anchor form equivalents of lower equating form one scores. It also produced the greatest anchor form equivalents of the higher equating form one scores. This result held, for the most part, across all of the forms of the GED tests studied.

For lower equating form one scores, the one-parameter method curve tended to be lower than the equipercentile curve which tended to be lower than the linear curve. The reverse appeared to hold for the higher scores. This relationship was present for most of the test forms studied.

The means of this bias, imprecision, and percentile comparison cross-validation statistics are presented in Table 3. Tukey critical differences are also presented.

Insert Table 3 About Here

Due to the differences in test lengths, none of these indices should be compared across tests.

The three parameter method produced the largest bias index for each test. The Reading Skills and Social Studies examinations were the

only tests for which the equating method F-test in the analysis of variance surpassed the .05 critical value. However, the difference in bias statistics among methods were not appreciably large.

The imprecision measure showed more substantial differences. The Tukey comparisons indicated that the three-parameter method was more imprecise than the other methods for all tests. No evidence of consistent differences among the other methods was found.

The percentile comparison measure for the three-parameter method was largest for each GED test. While none of the paired comparisons surpassed the Tukey critical difference, Scheffé comparisons of the three-parameter method with the mean of the equipercentile, linear, and one-parameter methods surpassed the .05 critical value (df = 3, 30) for Writing Skills (F = 3.52), Social Studies (F = 5.72), Reading Skills (F = 4.58), and Mathematics (F = 6.64).

Friedman Statistics (Conover, 1971) were calculated for each of the cross-validation indices and for each GED test because the assumption of normality was probably violated in the analyses. Forms were treated as blocks and equating methods as treatments in the analyses. The Friedman statistics were not reported here because the results were essentially equivalent to the analyses of variance.

Discussion

Factor Analyses

The factor analyses suggested that each of the GED tests, with the possible exception of the Mathematics test, were reasonable unidimensional. This was suggested despite the fact that the GED tests are, in general, content heterogeneous and contain many items which are presented

as part of a common stimulus block. Since consistent differences between Mathematics test equating results and those for other tests were not discovered, the results of dimenionality differences, if any differences did exist, were not detected by the procedures used. Unidimensionality may be crucial for item response theory scaling. For this reason, its assessment should be a routine aspect for any comparison among equating methods that includes item response theory methods.

Examination of Item Parameter Estimates

The three-parameter estimation procedure produced a number of extreme parameter estimates, suggesting that difficulties were encountered in parameter estimation. Since these difficulties can be expected to affect equating results, examination of item parameter estimates for extremes should be included in equating method comparisons. The existence of extreme unconstrained parameter estimates for item response theory models requiring constraints on other parameters to achieve convergence suggest difficulties in item parameter estimation.

Graphing of Equating Curves

The graphing and examination of equating curves suggested that a relationship existed among the equating curves. The relationships will be discussed later. Since relationships discovered can have consequences in practice, a graphing and examination of the equating curves can be very useful when comparing results from equating methods.

Some Factors Affecting the Cross-Validation

The findings from the cross-validation analyses necessarily depended on all factors affecting the adequacy of the equating. Equating

group sample size and the specific equating methods used are such factors. For example, different findings might have occurred had smoothing instead of linear interpolation been used in equipercentile equating or had Lord's (1980, pp. 209-210) suggested modification to LOGIST for extensive simultaneous estimation been used instead of the simultaneous procedure used here. Additionally, larger cross-validation samples can be expected to increase the chances of detecting differences among equating methods when differences do exist.

The cross-validation indices were designed to reflect differences between cross-validation anchor form and converted equating form (i.e., equated to the anchor form raw score scale) distributions for examinees taking both forms. Under equity considerations, the two distributions should be identical, apart from sampling error.

Bias Index

A bias index was calculated for each of the 55 test/equating form combinations. This index was the squared difference between the mean anchor form and mean converted equating form scores and reflects both equating error and error in sampling examinees for the cross-validation group. The three-parameter equating method tended to produce the largest bias indices in the cross-validation.

The larger bias indices for the three-parameter method may have reflected a combination of sampling error and imprecision rather than bias, as such, where bias is defined as the mean difference between anchor and converted equating form scores for an infinitely large cross-validation group. Consider the following not too unlikely scenario given a cross-validation sample of size 20. Suppose that there existed a single

examinee in the cross-validation sample with a very low score on the equating form. When converted to the anchor form scale using equating curves like those in Figure 1, a lower converted score would be produced for the three-parameter methods than for the other equating methods studied. If the sample happened to contain no very high scores to compensate for the very low one, then the bias index for the three-parameter method would be higher than the index for the other methods. Other likely scenarios with the same implications for the bias index are possible with small cross-validation samples. The chances of this phenomenon occurring should be minimized as cross-validation sample size increases.

The mean difference between anchor and converted equated form scores, over all 220 cross-validation examinees taking the anchor and equating forms of each test, was calculated to investigate this hypothesis. Although not presented here, the differences in means were smaller than might have been expected from the bias indices shown in Table 3. In fact, for the Mathematics test the mean difference for the three-parameter method was closer to zero than the mean difference for the other equating methods. Hence, the bias index may have been strongly influenced by the combination of imprecision and error in sampling the cross-validation group examinees.

The meaning of the larger bias indices for the three-parameter method is unclear. The bias index should be carefully interpreted when small cross-validation sample sizes are used.

Imprecision Index

A separate imprecision index was calculated for each of the 55 test/equating form combinations. This index represents the variance of the difference between cross-validation anchor form and converted equating

form combinations. The three-parameter equating method consistently produced the largest values of the imprecision index.

The imprecision index can be decomposed such that,

$$\frac{\Sigma(D_i - \bar{D})^2}{n} = s_x^2 + s_{y'}^2 - 2r_{xy'}s_x s_{y'}. \qquad (4)$$

The quantities s_x and $s_{y'}$ are the observed standard deviations, for examinees taking one of the equating forms, of the anchor form and converted equating form scores, respectively. The correlation between anchor form and converted equating form cross-validation scores is represented by $r_{xy'}$.

When comparing the imprecision index from one equating method to another, s_x^2 will remain constant. The quantities $s_{y'}^2$ and $r_{xy'}$ can vary, however. For most forms studied, $s_{y'}^2$ was largest for the three-parameter method. The quantity $r_{xy'}$ was not consistently larger or smaller for any of the methods. It appears that the larger imprecision indices for the three-parameter method resulted from comparatively larger variances of converted equating form scores for this method than for any other. Inspection of Figure 1 suggests how this occurred. Low equating form scores were converted to lower anchor form scores for the three-parameter method than for the other equating methods. High equating form scores were converted to higher anchor form scores for the three-parameter than for the other equating methods. This would be expected to lead to a larger variance of converted equating form scores for the three-parameter method and, therefore, a larger imprecision index.

A problem with the use of the imprecision index is amplified by considering equation (4). It can be shown that the use of linear regression, instead of linear equating, would be expected to lead to a smaller

value of the imprecision index. Although linear regression does not qualify as an equating method, since it does not meet the symmetry requirement, an equating method "could look better than it really was" if it produced too small a variance of converted equating form scores. The percentile comparison index was included as a potential procedure for circumventing this problem.

Percentile Comparison Index

The percentile comparison index was formed by calculating the difference between each integer anchor form score and the converted equating form score having an identical percentile rank in the cross-validation sample. Each difference was squared and weighted by the number of cross-validation examinees earning the corresponding anchor form score. The mean of these squared differences comprised the percentile comparison index. A separate index was calculated for each of the 55 test/equating form combinations.

It seems that this index will be larger whenever the variances of the anchor form and converted equating form cross-validation scores differ. (No proof can be offered as this is a fairly complicated index.) If this is true, then the problem of an equating method "looking better than it really was" which was mentioned in connection with the imprecision index would be eliminated with the percentile comparison index.

As with the imprecision index, the percentile comparison index tended to be largest for the three-parameter equating method. In both cases, the larger variances of the converted equating form distributions for this method were probably responsible for the larger values of the indices.

Critique of Cross-Validation Procedures

The cross-validation suggested that the three-parameter item response theory method produced inferior equating results. This was probably a result of the three-parameter method producing overly variable converted equating form score distibutions. Figure 1 illustrated how this probably occurred.

If the variances of the converted equating form scores of equating sample examinees (as opposed to cross-validation sample examinees) had been calculated, the variances for the three-parameter method probably would have been largest. (These variances are currently being examined by the authors.) Therefore, computation of means and variances of equating sample examinees have the capacity to provide useful information when completed prior to a cross-validation study.

The percentile comparison index was the only cross-validation index considered that is appropriate when randomly equivalent groups take the anchor and equating forms in the cross-validation. The bias and imprecision indices require that the cross-validation examinees take both forms. Also, the percentile comparison index is probably not biased in favor of a procedure like linear regression. However, it appears that the percentile comparison index is less sensitive to differences among equating method results since it less consistently identified differences than did the imprecision index. The percentile comparison along with the bias and imprecision indices should be used in equating method cross-validation studies.

As mentioned previously, the bias index can be affected by equating imprecision for small samples. Since the mean (bias) and variance (imprecision) of a distribution are generally not independent quantities, it may be beneficial to consider a composite index of the two, only.

Note that raw scores were used for all of the cross-validation comparisons. If the raw scores were linearly converted to standard scores so that each GED test had the same mean and standard deviation, then the indices could be compared across tests. We are currently attempting to use standard scores.

Implications for Test Equating

The problems encountered with the three-parameter equating method were at least partially a result of small sample sizes and the methods used to estimate the parameters. In review, the ability parameters and anchor form item parameters were estimated using over 2,000 examinee records for each test. The ability parameters were then fixed, as were the lower asymptote parameters. The difficulty and discrimination parameters were estimated, separately, for each of the equating forms using from 170 to 198 records for examinees taking the form.

The small sample sizes for the equating forms were probably responsible for the extreme parameter estimates discovered. Additionally, it was noticed that many examinee's scores on the anchor form and the other form of a test taken were very different. A screening procedure, for removing examinees whose scores on the two forms were very different, might have improved the situation. Lord's (1980, pp. 209-210) modification of LOGIST for extensive simultaneous estimation might also have improved the estimation.

A consistent relationship between the equating curves for the three-parameter method and other methods was found. An individual with a lower equating form one score would be penalized if the three-parameter equating curve, as opposed to the curve for any other method, was used

to convert the score to the anchor form scale. An individual with a higher equating form one score would have benefitted had the three-parameter equating curve, as opposed to any of the other curves, been used to transform the score to the anchor form scale.

This might have resulted from the problem encountered in the parameter estimation. However, in another study in which parameter estimation did not appear to be problematic, Kolen (1979), discovered a similar relationship when scores on a form were converted to the scale of another form that was slightly less difficult. Kolen (1981) hypothesized that this resulted from the condensing of the estimated true score scale with the three-parameter model. That is, estimated true scores below the "pseudo-chance" level of the test (the sum of the lower asymptote parameters) do not exist. Hence, the estimated true score scale is a condensed version of the raw score scale. This hypothesis takes on greater weight when it is realized that the equating curves seem to pass at or very near the joint raw score means of the equated forms. The condensing problem can be avoided if estimated observed score equating (Lord, 1980) were used instead of the estimated true score equating used here.

Any differences among the one-parameter (Rasch), equipercentile, and linear methods which might have existed were not detected by the cross-validation procedures. The use of the larger cross-validation groups would be expected to lead to discovery of these types of differences, if they did exist.

Conclusions

The following procedures were found to provide useful information in a comparison of equating methods and should be considered for use in future studies.

1. Tetrachoric factor analyses to assess the degree of test unidimensionality.
2. Examination of graphs of equating curves to detect idiosyncratic results.
3. Examination of item parameter estimates for extreme estimates.
4. Comparison of the equating sample means and variances of anchor and converted equating form scores.
5. Completion of a cross-validation study including the calculation of bias, imprecision, and percentile comparison indices.

The cross-validation analyses require the existence of a representative group of examinees who were administered both equated forms but were not included in the equating. The percentile comparison index can still be used when randomly equivalent groups of cross-validation examinees take the forms.

Larger equating sample sizes and/or a modification of the estimation procedure would be necessary before the three-parameter method could be suggested for equating the GED tests. The other three equating methods produced similar cross-validation results. The one-parameter (Rasch) equating method can be expected to produce results which are as stable as those for linear and equipercentile methods for equating achievement tests which are similar to the GED with sample sizes around 200 and when both forms of similar difficulty have been administered to the same examinees. The three-parameter equating method is much less stable in this situation. Additionally, investigation of the possibility of score scale condensing with the three-parameter method is warranted.

References

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Conover, W. S. Practical non-parametric statistics. New York: Wiley, 1971.

ETS. The final report for a project to develop twelve new forms of the Tests of General Educational Development and to standardize the tests nationally in the United States. Princeton, N.J.: Educational Testing Service, 1978.

GED teacher's manual for use with official GED practice tests. Washington, D.C.: GED Testing Service of the American Council on Education, 1979.

Kolen, M. J. Comparisons of equipercentile, linear and selected latent trait methods for equating forms and levels of the seventh edition of the Iowa Tests of Educational Development. Unpublished Ph.D. Dissertation, The University of Iowa, 1979.

Kolen, M. J. Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement. In press.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N.J.: Erlbaum, 1980.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Wright, B. D. & Stone, M. H. Best test design: A handbook for Rasch measurement. Chicago: MESA, 1979.

Table 1

First 12 Eigenvalues and Percentage of Variance Accounted for in Tetrachoric Factor Analysis of Anchor Forms

Factor	GED Test									
	Writing Skills		Social Studies		Science		Reading Skills		Mathematics	
	Eigen-Value	% Variance	Eigen-Value	% Variance	Eigen-Value	% Variance	Eigen-Value	% Variance	Eigen-Value	% Variance
1	20.79	25.98	18.66	31.10	17.09	20.48	14.56	36.40	15.07	30.14
2	2.80	3.50	1.83	3.05	2.30	3.83	1.57	3.92	3.22	6.40
3	2.07	2.58	1.39	2.32	1.87	3.17	1.30	3.25	1.48	2.96
4	2.00	2.50	1.34	2.23	1.46	2.43	1.16	2.90	1.40	2.80
5	1.74	2.18	1.25	2.08	1.39	2.32	1.05	2.63	1.21	2.42
6	1.43	1.79	1.19	1.98	1.28	2.13	1.01	2.52	1.16	2.32
7	1.38	1.72	1.16	1.93	1.22	2.03	1.00	2.50	1.11	2.22
8	1.34	1.68	1.08	1.80	1.14	1.90	0.98	2.45	1.05	2.10
9	1.29	1.61	1.07	1.78	1.12	1.87	0.93	2.32	1.04	2.08
10	1.26	1.58	1.07	1.78	1.08	1.80	0.87	2.18	1.02	2.04
11	1.18	1.48	1.04	1.73	1.05	1.75	0.87	2.18	0.98	1.96
12	1.16	1.45	0.98	1.63	1.03	1.72	0.84	2.10	0.95	1.90

Table 2

Standard Deviation of Item Response Theory Difficulty and Discrimination Parameter Estimates

Test	Form[1]	Number of Items	Parameter Estimate		
			One Parameter Difficulty	Three Parameter Difficulty	Three Parameter Discrimination
Writing Skills	Anchor	80	0.55	0.95	0.27
	Equating	880	0.52	2.99	0.32
Social Studies	Anchor	60	0.50	0.90	0.31
	Equating	660	0.53	6.42	0.31
Science	Anchor	60	0.62	1.38	0.42
	Equating	660	0.53	7.98	0.30
Reading Skills	Anchor	40	0.61	0.91	0.38
	Equating	440	0.57	1.18	0.29
Mathematics	Anchor	40	0.81	1.28	0.45
	Equating	440	0.78	1.61	0.36

[1]Form Equating refers to equating forms one through eleven taken together.

Table 3

Mean Cross-Validation Indices and Tukey Critical Differences

Index	Equating Method	GED Test				
		Writing Skills	Social Studies	Science	Reading Skills	Mathematics
Bias						
	Equipercentile	3.14	1.27	5.02	0.63	1.03
	Linear	2.98	1.30	4.82	0.61	1.11
	One Parameter	3.24	1.28	4.38	0.63	1.13
	Three Parameter	3.85	1.70	5.28	1.04	1.55
	Tukey Critical Difference	1.05	0.51*	2.22	0.32*	0.70
Imprecision						
	Equipercentile	63.02	48.85	48.10	22.36	28.76
	Linear	61.81	47.93	47.12	21.93	28.77
	One Parameter	66.25	50.23	49.75	22.26	31.85
	Three Parameter	73.47	60.41	57.21	28.62	36.74
	Tukey Critical Difference	5.60*	8.21*	5.24*	3.75*	3.79*
Percentile Comparison						
	Equipercentile	17.07	15.05	13.99	7.69	7.99
	Linear	15.81	14.96	12.35	7.26	7.34
	One Parameter	16.31	14.67	12.02	7.18	8.34
	Three Parameter	20.20	20.72	15.34	10.15	11.32
	Tukey Critical Difference	4.51*	5.42*	4.53	2.95*	2.96*

*The equating methods main effect surpassed the .05 level of significance in the analysis of variance.

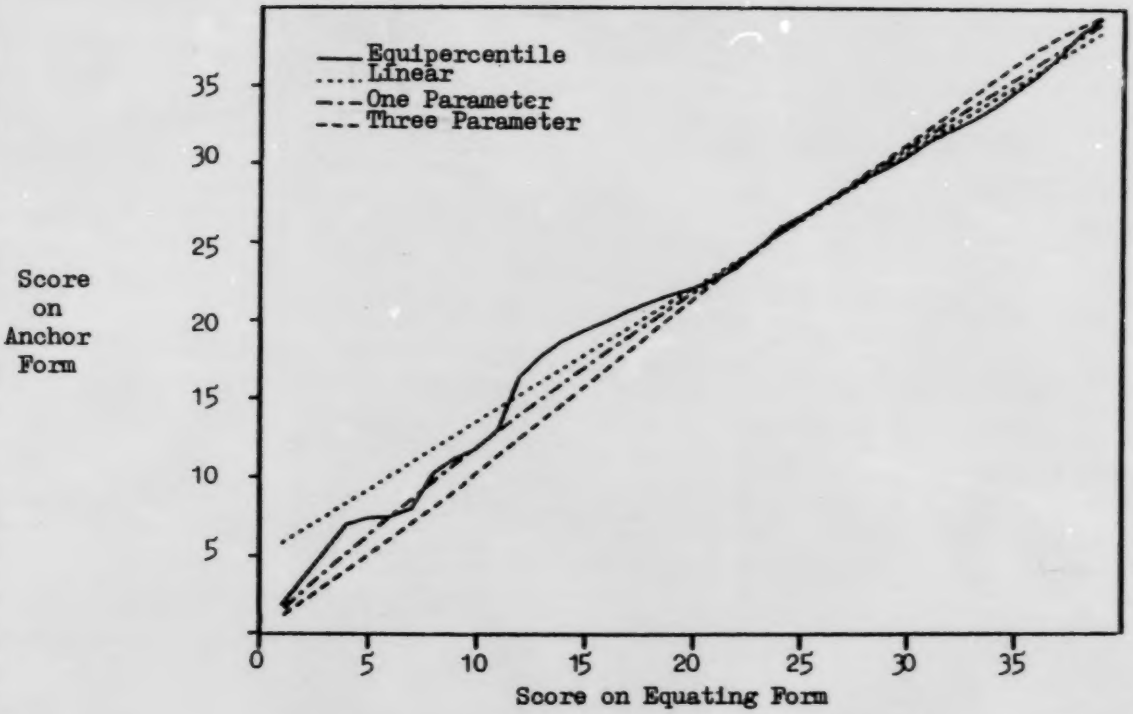

Figure 1. Equating relationships between the anchor form and equating form one of the GED Reading Skills test for four equating methods.